DOCUMENT RESUME

ED 235 237 TM 830 664

AUTHOR Green, Donald Ross
TITLE Content Validity of Standardized Achievement Tests and Test Curriculum Overlap.
PUB DATE Apr 83
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Curriculum; Curriculum Development; Curriculum Evaluation; Elementary Secondary Education; *Standardized Tests; Test Construction; *Testing Problems; *Test Validity
IDENTIFIERS *Test Curriculum Overlap

ABSTRACT

Discussions of "test curriculum overlap" that focus on the term "mismatch" tend to be based on several untenable premises. This paper addresses the issue of the degree to which standardized tests should reflect the specific content of a given school curriculum with regard to three points: (1) The idea of matching the content of standardized tests to school curriculum is based on the misperception that a specific standard curriculum can be defined. There can be great variability between schools and even within the same school. (2) Content validity is relative to the purpose for which the data are used. The content specifications appropriate for a teacher-made test designed to assess immediate outcome of instruction are not appropriate for standardized tests designed to assess year-to-year progress and to establish unambiguous norms. (3) To serve the purposes of curriculum planning and evaluation, the procedures used by publishers to establish content validity must provide coherence and continuity over a range of grades. The content must be a composite of material related to common threads and principles inherent in subject fields. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED235237

Content Validity of Standardized Achievement Tests and Test Curriculum Overlap

Donald Ross Green

CTB/McGraw-Hill

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. R. Green

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, 1983

TM 830 669

PRINTED IN USA

Abstract

Discussions of "test curriculum overlap" that focus on the term "mismatch" tend to be based on several untenable premises. This paper addresses the issue of the degree to which standardized tests should reflect the specific content of a given school curriculum with regard to three points.

1. The idea of matching the content of standardized tests to school curriculum is based on the misperception that a specific standard curriculum can be defined. There can be great variability between schools and even within the same school.

2. Content validity is relative to the purpose for which the data are used. The content specifications appropriate for a teacher made test designed to assess immediate outcome of instruction are not appropriate for standardized tests designed to assess year to year progress and to establish unambiguous norms.

3. To serve the purposes of curriculum planning and evaluation, the procedures used by publishers to establish content validity must provide coherence and continuity over a range of grades. The content must be a composite of material related to common threads and principles inherent in subject fields.

Introduction

It has been common practice for many years to say that schools should select standardized tests that fit their curriculum (e.g., Ruch, 1929; Remmers, Gage, & Rummel, 1965). In recent years this point of view has been modified by many to say instead that schools should only use standardized tests that match exactly what is taught, item by item, no more no less (Freeman, et. a., 1979; Tyler & White, 1979). Anything else is held to be "mismatch". Please note that mismatch is a term designed to lead one immediately to the conclusion that something is very wrong. Consequently, I prefer to use the more neutral phrase "test curriculum overlap."

This paper will present a view contrary to that of the mismatch theorists holding instead that something substantially less than 100% overlap is 1) an absolute necessity and 2) highly desirable. The reasons for these assertions are many. I will limit myself to three of them because:

a. I don't wish to repeat what I said last year (Green, 1982).

b. These three suffice to make the point

c. Mehrens (1983) will present other statements at least as cogent and in a more compelling style than I can manage.

d. These will take all the time I have anyway.

My three arguments are based on:

1. What one finds when examining the question "What is the curriculum?". This will show that 100% overlap is not practical and suggest why it is not desireable.
2. The nature of uses and users of standardized achievement tests. This information further supports the point about the desirability of a less than perfect match.
3. The procedures for establishing content validity typically used by publishers, i.e., the test construction process. The first two arguments explain why these procedures are appropriate and these procedures in turn show again that less than 100% overlap is both necessary and desirable.

It should be understood that the tests being discussed in this paper are those found in the major achievement batteries in current use around the country. Probably almost all students in American schools take a battery of tests of this sort several times in the course of their school careers and a number of them take a standardized achievement test one or more times each year.

However, it should be borne in mind as well that these students take other tests, namely teacher made tests, far more frequently. In most cases teacher made tests are designed to test precisely what has been taught by that teacher in that school to that class in order to evaluate student learning of the part of the content of the curriculum just presented. In

such tests for that use the unintentional inclusion of material which does not relate to the portion of the curriculum just studied may properly be called a mismatch. Where it occurs deliberately because the teacher is trying to get at something else as well, it is a matter to be explained to the students who seldom think it fair even though the teacher's reasons may be good. It is probably because so many people experienced these feelings as students that the notion of mismatch is so readily and uncritically accepted.

Mismatch theory appears to confuse standardized achievement tests with teacher made tests. However, standards for the content validity of the tests teachers make to assess immediate student learning are not appropriate for standardized achievement tests. The content specifications of the two differ as do the construction procedures, and they do not serve the same purposes. To understand these assertions it will be helpful to consider first the question "What is the Curriculum?".

What is the Curriculum?

This is not meant to be a deep and difficult question although one can easily make it so. It is merely being used to point to the great variability found within and between schools of the experiences, planned or not, that students have in any curricular area. "The curriculum" is not a monolithic entity perceived in the same way by all. Yet if a test is to match a curriculum it is plain that the curriculum must be specified. The question is what version of the curriclum is to be used? The possiblities are several. The list might look like this:

Curriculum A. The curriculum perceived by the student.

Curriculum B. The curriculum planned by the teacher for that student or for the class as a whole.

Curriculum C. The curriculum specified by the teacher's supervisor (e.g., department chairman, principal, curriculum supervisor) -- there could be several of these.

Curriculum D. The curriculum as represented by the school system syllabus or list of goals and objectives -- if it exists.

Curriculum E: The curriculum as represented by the state curriculum guidelines -- if they exist.

Curriculum F: The national consensus curriculum -- to the degree that such a thing exists. Obviously there is no official national curriculum, but nationally standardized tests are attempts at this.

At some level of generality all these curricula will have much in common. But note that no course of instruction can cover -- much less get students to learn -- everything any more than any test can cover everything. Both end up being samples of some domain. Thus choices must be made for a test.

Students may well feel that to be fair any tests they take should exactly match Curriculum A and they may consider one that matches only Curriculum B something of a mismatch; few teachers would agree with this. Similarly teachers sometimes say that the tests should match Curriculum B, as their own usually do, otherwise they will be

misjudged; few curriculum supervisors take the results of teacher made tests as appropriate data for program evaluation or curriculum planning.

Curriculum personnel may in turn want a closer match to their version of the curriculum than do the superintendent and school board. I recall one mathematics supervisor who was very upset when the system insisted on using a test which had items that included fractions, especially because the result was that the system decided that work on fractions was needed. The supervisor felt the test was a mismatch (see Taylor, 1979). Yet the school board felt they had been provided necessary information. A perfect match to the supervisor's official curriculum would have concealed information the school board valued.

Clearly perfect matches are not possible from all points of view. Thus teachers' tests can fit the curriculum they believe they have presented but that will not be the curriculum perceived by all the students. A school system test may match the official curriculum of that system better than any published test but the match to the curriculum presented to some classes may be pretty poor while the match to the policy makers expectations may be worse if they ever discover what the students were not taught.

On the other hand, if the students have learned fundamental skills and knowledge and understand it, they will be able to answer many questions dealing with material not directly taught. It

is true that teaching something directly is more effective than relying on transfer and generalization but it is also true that generalized skills and understandings do develop albeit not quickly. Since all the specifics can never be taught in most subjects, this development is highly desirable and tests of long term progress should try to assess it. This can only be done by having items that ask about content not directly taught.

In short, there are many versions of the curriculum and 100% overlap between tests and all of them is logically impossible. However, giving tests that include material not in the curriculum as defined by one set of people often turns out to be very useful for those wishing to monitor the work of this set of people. A description of who uses standardized tests for what may make this point clearer.

Uses and Users of Standardized Achievement Test Batteries

Standardized achievement test batteries are used for a large variety of purposes by a large variety of people. The purposes may be grouped into three overlapping categories. These are accountability, program planning and evaluation, and instructional planning and evaluation. The users range from legislatures and school boards on the one hand to teachers, students, and parents on the other.

Members of the first set of groups are policy makers who tend to be most concerned about accountability and therefore are primarily interested in large scale group data; test data are only one of the sources of information used for this sort of purpose but they are an important one.

Nevertheless, these policy makers rarely examine the tests themselves and spend little time examining the data. Instead they usually rely on professional staff personnel to select the tests, judge their validity and provide interpretations of the data.

At the other end of the user spectrum it is the teachers who interpret the data. They quite naturally focus on instructional planning and evaluation and are interested largely in the individual student data; standardized test data are also only one of their sources of information for these purposes but in this instance these data do not usually play a very important role (Lazar-Morrison, et. al, 1980). However, important or not, teachers do not tend to spend much time considering such data nor do they do so very often. After all, the tests are not given frequently.

This does not mean teachers are not concerned about validity; they are concerned about content validity which they tend to see in terms consonant with mismatch theory partly because of fears of being judged or misjudged on their effectiveness. Typically teachers mediate on the reporting of the results of standardized tests to students and parents and the data have impact on them only if the teachers so choose. As already noted, teachers do not pay much attention to test data so that is probably rare.

Nevertheless, there is a group who not only find the information yielded by standardized tests important but who use it often. These users are the professional personnel who fall in between the two groups just discussed; they are the curriculum supervisors and planners, the directors of testing and evaluation, the

managers of special programs, the school principals, the school counselors and any number of other professionals; collectively these people use standardized test data for all three categories of purposes. Furthermore, the data are used often and are considered important.

It is this group of middle administrators and professionals that is usually the most influential in decisions to purchase standardized tests. They know about and care about the technical qualities and features of the tests as well as their content validity and they provide the most information to publishers about what these tests should be like to serve the information needs of the district. In most cases this means that answering policy makers questions about general status and year to year progress comes first with program planning and evaluation second. Instructional evaluation is usually dead last.

To be sure, especially in the larger districts, test batteries are selected by committees which include a number of teachers and sometimes these teachers focus on the match of the content of the tests to their own teaching because of their fears of being unfairly evaluated. But even though they try hard to make these tests useful to teachers publishers cannot really respond to such concerns except to also offer criterion referenced tests.

The publishers of any test batteries meant to be used on a large scale, such as a district wide program, will try first and foremost to have them yield information that will help the middle managers and professionals meet their responsibilities. Particularly when their

responsibilities call for reports to the district policy makers about school and district status and change. The test must be such that it can have meaningful national norms. As I noted last year (Green, 1982) the requirement of national norms means that a test designed to fit a local curriculum will not do. At this point it may be well to consider just how publishers proceed to build standardized norm referenced achievement tests that can validly serve these functions.

Content Validity of Standardized Achievement Tests

The content validity of an achievement test is an inference derived from an examination of the procedures used to determine the content and nature of the items that end up being included in the test. Content validity cannot be determined solely by examining those items and seeing how many of them belong in certain categories. Perhaps a description of the test construction procedures will make the reason for this last statement plain. This description is an abstraction of what has been general practice at CTB for the last dozen years. It is believed that these procedures do not differ dramatically from those used by other publishers (cf. Burrill).

The first step in planning an achievement battery is to decide into what general domains the content will be divided. Some of these are well defined and essentially universal (e.g., reading comprehension) and some are not (e.g., study skills), but in any case each leads to a separate test or hierarchical

series of tests whose validity must be considered separately. Also because the batteries are intended to be used in large scale assessment programs, only domains that can be measured by paper and pencil objective tests are chosen. The test battery is most commonly a revision of a battery already published and so while changes are often made, their basic composition is already established. In any case the matter of choosing the general domains to be included in a battery is not really a concern here because it is within these areas that the issues considered in this paper arise rather than with their choice.

If the publisher is not just entering the field, the beginning point is an analysis of the latest tests on their current list and especially the prior edition of the test in question. A parallel analysis of the latest available curriculum guides and the most widely used textbooks, especially basal series where they exist, provides the basis for planning changes. The guides used are those developed by states and by the large school districts. There are many of these and they vary in recency, in quality, and in the degree to which they are followed and used in the schools. Consequently choices must be made as to which ones to consider. The textbooks are also numerous and also vary in recency and in quality. Again choices must be made. A consideration in both these choices is the fact that the materials reviewed reflect the past and the task is to predict what school curricula will be like two to four years down the road when the test is published. The opinions of curriculum experts, both academic and those in school systems, are brought to bear here as well

as at a number of other points during the process.

Next the test designer's task is to abstract the generally common elements, perhaps in the form of instructional objectives but usually including some dimension referring to level of complexity and process as well. However it is done, it is necesary to put them into some coherent form. This is a critical task. It is this document that defines the content domain to be represented in the test in its tryout form. This means that elements may subsequently be deleted but that nothing can be added. Coherence and continuity over a range of grades are the important attributes; even if it is written as a set of objectives it cannot be merely a list. It must have the organizing elements that local curriculum guides tend to lack (Goodlad, 1983).

From this domain the editors select the specific objectives to be included in the tests. This is also a critical task. Obviously this selection can be only a sample of the objectives in the domain simply because with a few exceptions in the primary grades, such as the names of letters, the population is either not finite or it is very large. The sample of objectives chosen for inclusion usually cannot be very large relative to the size of the domain because there are obvious practical limits on test length. The objectives chosen must include all those judged to be critical as well as sample the range of the domain as fully as possible. To do these two things well is a difficult art.

A few of the difficulties may be mentioned. First, the content covered by any one level of the test must cover a broad enough range to reflect growth over a span of at least one and often several

grades. If the span is made too broad the test will yield less specific information and will be somewhat insensitive to growth. If the test is too narrow, too many students in the same grade and even the same class will hit its floor or ceiling.

Second the final selection of content must not only have some abstract coherence that will satisfy the curriculum experts in the field, it must have a strong relationship to what students actually learn. Only if this condition is met will the resulting test scores have the necessary characteristics. For example, a) the scores must display a reasonable distribution when the test has been taken by groups at the level for which the test is designed, b) students who have more instruction than others should score higher on the average, and c) the test should display a substantial degree of internal consistency. If the test contains much material that is not related in the students' heads these results will not be obtained. It is important to recognize that this requirement is not the equivalent of a match to curriculum guides or to textbook content; just because something appears in the instructional material and is taught does not mean that it becomes a functional part of student knowledge.

For example, in developing a criterion-referenced test in mathematics CTB editors found that some objectives were taught and partly learned in grade six by some students but that the students in the two subsequent grades did less well on these items; however, after these grades performance on the objective

increased substantially. It is noteworthy that all of the more than a dozen items written to each of these objectives showed the same behavior. Attempts to calibrate these items along with those from CTBS/U failed because, unlike most of the material, they did not relate consistently to the CTBS trait.

Perhaps this was merely material that was not practiced or retaught during the period but not all skills are forgotten if not retaught. It all depends on the degree to which the concepts or skills are related conceptually and functionally to what the students are learning during the period. If the material becomes a part of their structure of understanding or their operational skills student performance is not likely to decline. If the material is independent of such skills it will tend to be forgotten by many.

Objectives which do not show growth through the grades can be included in criterion-referenced tests but not in a test which must provide unambiguous norms. The curriculum may say this content is part of the course of instruction but the evidence says that it is a peripheral part, not an integral part.

The remaining steps of the test construction process are long and arduous but need to be described only cursorily since the principal points have been made. Once the objectives to be included have been chosen the next step is to develop specifications for the items to measure these objectives. As with the objectives the items written are in most cases only a sample of the population of items that could be written to the objective. The items are written, edited, reviewed, tried out, rewritten, reedited, reviewed again. Many

drop by the wayside but eventually a set of items believed to provide the most adequate representation of the domain originally specified is assembled. This version is usually the standardization edition for which norms are obtained.

In sum, standardized achievement tests are designed to measure growth over substantial periods of time such as fall to spring or year to year. Short term learning (e.g., week to week or month to month) will not be readily detected by these tests unless it is both fully integrated in to the basic trait (an event that may not occur upon initial instruction because overlearning is often needed) and represents a large change. These characteristics are necessary if the test is to have meaningful norms and will be used in more than one grade.

It will be easily seen that the content of a standardized norm-referenced achievement test is not, cannot, and was never meant to be a pure match to the curriculum of any school system, school or class. It should be clear that the content is a very carefully selected composite of the views of many people modified as necessary by the results of tryouts.

The resulting test will largely contain those elements of the curriculum which are common to many schools and school systems. It will emphasize those aspects of achievement that show growth over substantial periods of time and that represent coherent bodies of knowledge and skills. Such a test can serve many purposes, notably a large portion of those discussed in the previous section.

Summary and Conclusions

Discussions of test curriculum overlap that focus on the term mismatch tend to be based upon several untenable premises.

In the first place the assumption is made that the curriculum presented to and experienced by the student is what should be tested. It is argued in this paper that this is impossible because there are so many versions of the curriculum so defined. It is also argued that it is better to define the domain to be measured as broadly as one can and still retain coherence. The advantages of this include learning about achievement in aspects of the domain not necessarily contained in any one version and being able to compare the outcomes of different versions of the curriculum.

Second, it is assumed that the purpose of standardized testing is the evaluation of the immediate outcomes of instruction objective by objective. A corollary is that teachers are the primary users of standardized test data. This assumption confuses the functions of teacher made, or end of chapter, or criterion referenced tests with those of standardized tests which have many other uses and many other users. In particular they are used to assess long term growth in programs, schools, and school systems.

Third, it is assumed that items written to any nominally related set of objectives or topics can be combined into a test which can be normed and be useful for system wide use. This is

not true. Not all the material found in textbooks and in local curricula relate to the common threads and principles inherent in subject fields. Only the latter belong in tests which will be used in the ways that standardized tests are used.

Finally, I would not want anyone to think that what I have just said was derived from profound new insights into test validity so I will close with a quote from the past. Thirty years ago in his well known chapter on validity, Cureton (1951) wrote about this topic. He noted that the point of view that tests and curricula should match very closely is based on the assumption that "... there is no need to inquire into the relevance of an immediate objective to the ultimate objectives of education" (Cureton, 1951 p. 669). As did he, I believe that there is such a need and that standardized norm-referenced achievement tests provide a means for doing this much more adequately than tests which try to represent all the elements of a particular instructional program. "As we move from ordinary unit and course and subject achievement tests to tests covering larger areas of knowledge and behavior, and from these to still more comprehensive tests which attempt to assess the generality of students' educational development, the value of curricular relevance becomes progressively lower" (Cureton, 1951 p. 671).

In short, a very high degree of test curriculum overlap is an overrated goal and the term mismatch in this context is misleading.

References

Burrill, L. E. How a standardized achievement test is built. Test Service Notebook 125, Harcourt Brace Jovanovich, Inc.

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1951.

Freeman, D. J. et. al, A closer look at standardized tests. Research Series No. 53, East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, 1979.

Goodlad, J. I. Individuality, commonality, and curricular practice. In G. D. Fenstermacher and J. I. Goodlad (Eds.), Individual differences and the common curriculum (Part I). Chicago, Illinois: The National Society for the Study of Education, 1983.

Green, D. R. Matching tests to curricula and test equating. Paper presented at the annual meeting of the American Educational Research Association, New York City, March 1982.

Lazar-Morrison, C., et. al., A review of the literature on test use. CSE Report No. 144. Los Angeles, California: Center for the Study of Evaluation, 1980.

Mehrens, W. A. National versus local tests and curriculum: Inferences to which domain and why and implications of any mismatch. Paper presented at a Symposium at the annual meeting of the National Council on Measurement in Education, Montreal, April 1983.

Remmers, H. H., Gage, N. L. and Rummel, J. F. A practical introduction to measurement and evaluation (2nd ed.). NY: Harper & Brothers, 1965.

Ruch, G. M. The objective or new-type examination. NY: Scott, Foresman, 1929.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-147.

Taylor, R. Mathematics: A view from the schools. In R. W. Tyler and S. H. White (Eds.), Testing, teaching and learning. Report of a conference on research on testing. Washington, D.C.: Department of Health, Education and Welfare. National Institute of Education, 1979.